Talanta 83 (2011) 1209–1224

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Data processing pipelines for comprehensive profiling of proteomics samples by label-free LC–MS for biomarker discovery

Christin Christin[a,b], Rainer Bischoff[a,b], Péter Horvatovich[a,b,*]

[a] Analytical Biochemistry, Department of Pharmacy, University of Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands
[b] Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Available online 11 November 2010

Keywords:
Computational biology
Quantitative data processing
Statistical analysis and validation
Label-free quantification
Biomarker discovery
Comparative LC–MS profiling

ABSTRACT

Label-free quantitative LC–MS profiling of complex body fluids has become an important analytical tool for biomarker and biological knowledge discovery in the past decade. Accurate processing, statistical analysis and validation of acquired data diversified by the different types of mass spectrometers, mass spectrometer parameter settings and applied sample preparation steps are essential to answer complex life science research questions and understand the molecular mechanism of disease onset and developments. This review provides insight into the main modules of label-free data processing pipelines with statistical analysis and validation and discusses recent developments. Special emphasis is devoted to quality control methods, performance assessment of complete workflows and algorithms of individual modules. Finally, the review discusses the current state and trends in high throughput data processing and analysis solutions for users with little bioinformatics knowledge.

1. Introduction

The recent widespread application of mass spectrometry to quantify and identify large numbers of compounds in biological matrices leads to an explosion of acquired data. The goals of these measurements are to explore the underlying molecular mechanism of disease, to identify compounds (biomarkers) strongly related to the stage of the disease, its onset or progression for diagnostic purposes, to identify novel drug targets, and to follow the efficiency of treatment. The dynamic behavior of multifactorial diseases requires a systems biology approach to find reliable biomarkers taking molecular regulatory mechanisms, compound flux and concentration changes into account [1]. To explore robust changes in molecular systems related to disease, it is necessary to analyze a large number of samples from different biological entities, for example from different, clinically well characterized patient groups. Generally biomarker research is based on complex biological samples containing a large number of diverse compounds such as proteins, peptides and metabolites. Liquid chromatography coupled to mass spectrometry (LC–MS) is one of the most widely used comprehensive profiling techniques to measure compounds in biological materials. A single comprehensive LC–MS analysis cannot cover all types of compounds in the samples. Instead, it measures one class of compounds such as metabolites, lipids, and proteins leading to biomarker discovery in this class of molecules. Even with a technique targeting one of the above mentioned classes of compounds, not all types of molecules can be measured due to ionization limitations of the electrospray interface. Another challenging problem is the wide dynamic concentration range of the compounds, which can reach 9–11 orders of magnitude in the case of body fluids such as blood [2,3]. From this wide dynamic concentration range, modern mass spectrometers are only able to cover 2–4 orders of magnitude. The gap between the existing and

Abbreviations: 2D-LC–MS, two dimensional liquid chromatography coupled to mass spectrometry; AMT, accurate mass and retention time tag; APEX, absolute protein expression; APML, annotated putative peptide markup language; Cap-LC–MS, LC–MS system equipped with capillary LC column (1 mm internal diameter) and using ionspray for ionization; Chip LC, LC–MS system equipped with nano-LC column (75 μm internal diameter) integrated in a microfluidic device and using electrospray for ionization; CID, collision induced dissociation; DDA, data dependent acquisition; emPAI, exponentially modified protein abundance; ETD, electron transfer dissociation; FTMS, Fourier transform ion cyclotron resonance mass spectrometry; HDSS, high dimensionality small sample size problem; HUPO PSI, Human Proteome Organization Proteomics Standard Initiative; LC, liquid chromatography; LC–MS, liquid chromatography coupled to mass spectrometry; MEND, matched filtration with experimental noise determination; MRM, multiple reaction monitoring; MS, mass spectrometry; MS/MS, fragment ion mass spectra of selected precursor ions; MS-1, single stage mass spectrometry; nano-LC, liquid chromatography using chromatographic column of internal diameter smaller than 100 μm; PTM, post translational modification; RSD, relative standard deviation; XML, extensible markup language.

[*] Corresponding author at: Analytical Biochemistry, Department of Pharmacy, University of Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands. Tel.: +31 50 363 3341; fax: +31 50 363 7582.
E-mail address: p.l.horvatovich@rug.nl (P. Horvatovich).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.10.029

measurable dynamic concentration range can be reduced by using comprehensive fractionation (4–6 orders of magnitude), multidimensional chromatography (up to 8 orders of magnitude) [4] or targeting a specific subclass of compounds, *e.g.* by using an affinity enrichment step of a certain type of glycoproteins on a lectin column (up to 5–7 orders of magnitude) [5]. Another challenging factor is that although proteins and protein complexes are directly involved in the molecular processes of biological phenomena, their peptide constituents obtained after enzymatic cleavage are measured since they are more suitable for liquid chromatography analysis and have better ionization properties than intact proteins or protein complexes. The most widely used endopeptidases cut proteins at well-defined sequence positions, resulting in non-overlapping peptides mixtures, from which only a fraction of theoretical possible peptides are detected. In this peptide centric approach also called as “bottom-up”, or “shotgun” strategy, the quantity of initial proteins is determined indirectly based on few or more peptides, which leads to misleading quantification and identification in the presence of multiple highly homologous proteins having one or few peptides in common, proteins with multiple splice variants, proteins presenting different degrees of post-translation modifications (PTMs) or in the presence of various truncated forms of the same protein [6,7].

Biomarker discovery requires close collaboration between medical researchers, analytical chemists and bioinformaticians in order to obtain the relevant molecular information related to different aspects of disease [8,9]. This includes patient cohort selection, sampling of the biological material, sample storage, sample preparation, choice and optimization of LC–MS profiling platform, data analysis providing protein identifications, quantification, statistical analysis and experimental validation of the results. Several review papers describe the various techniques and steps of the protein profiling for biomarker discovery in detail [9,10].

Bioinformatics plays an important role in this process as it has the goal to extract quantitative and qualitative information for a large number of compounds (proteins and metabolites) that are present in complex biological samples and to select the discriminatory compounds between predefined sample sets. Recent advances in sample preparation methods, liquid chromatography and mass spectrometry instrumentation resulted in a large diversity of acquired data. This results in a huge challenge for bioinformatics to provide reliable information extraction and knowledge generation approaches. The computational tools must evolve continuously to keep up with the different types of generated data. Besides direct information extraction and knowledge discovery from raw data, bioinformatics plays an important role in experimental design, quality assessment of the profiling platform, sampling methods, sample handling, storage and preparation methods, or quality control of data pre-processing, statistical analysis and statistical validation.

This review focuses on fundamental data processing and current challenges in supporting biomarker discovery research in proteomics for diagnosis and treatment follow-up using LC–MS of label-free, shotgun proteomics data, highlighting significant innovations in the bioinformatics field such as new algorithms, data integration, high throughput automatic data preprocessing solutions, quality control of different data processing modules and complete workflows, including assessment of the quality of sample preparation steps and LC–MS profiling platforms [9,11–19]. We will also investigate how insights from analytical chemistry contribute to parameter optimization leading to the development of novel bioinformatics applications that provide more accurate and reliable information extraction from the raw data. Alternative approaches based on differential labeling of samples with reagents having the same chemical but different stable isotope constitution have been covered in other reviews [20–27] and will not be treated here. This review limits the discussion further to biomarker discovery aiming to determine comprehensively the identity and quantity of sample constituting proteins using analytical methods with low sample throughput. Biomarker validation using analytical methods with high sample throughput providing quantitative information on preselected list of proteins by using analytical methods such as multiple reaction monitoring, antibody arrays and ELISA will not covered here. Recommendation on analytical, clinical and informatics aspects of biomarker discovery and validation as well their limitations was discussed recently by several reviews [28–34].

2. Data processing pipelines in LC–MS

LC–MS has become the major platform for analyzing samples in biomarker discovery research due to its relatively high throughput (60–90 min for analysis of one sample), sensitivity, selectivity and coverage of many peptides and proteins [9,35,36]. In label-free LC–MS experiments, proteins or produced tryptic peptides are not modified chemically and their isotope constitution is unchanged. In label-free experiments, a large number of samples are analyzed independently by LC–MS resulting in corresponding raw data files. The quantitative and compound identity information is extracted using dedicated data processing pipelines. This is followed by matching compound quantity and identity across several chromatograms resulting in a matrix containing quantitative information about a large number of compounds in the different samples. In shotgun proteomics approach the target compounds are proteins, therefore methods are required to determine the original protein composition of samples and their quantities based on incomplete set of measured constituting peptides. Compounds discriminating between predefined classes of samples are obtained from this matrix using dedicated statistical analysis and validation pipelines. When a systems biology approach is involved in the biomarker discovery process, it is necessary to couple the list of discriminating proteins to protein interaction (*e.g.* STRING, BIND) or pathway (*e.g.* KEGG) databases [21,37] to elucidate the disease mechanism. Fig. 1 shows the main parts of a generic proteomics pipeline for biomarker discovery.

Most of the signals measured by LC–MS are not related to real compounds but are part of white noise, background ions or simply chemical noise. Different mass analyzers generate data of different structure due to differences in scanning speed, mass resolution, measured dynamic concentration range, changes in peak width and resolution across the m/z domain and varying mass accuracy [38]. The most common mass analyzers applied in proteomics biomarker research are quadrupole, 3-dimensional quadrupole iontrap, 2-dimensional linear iontrap, time of flight, and inductively-coupled resonance (ICR) trap family of mass spectrometers such as Orbitrap and Fourier transform ion cyclotron resonance mass spectrometers (FTMS) [39]. Besides mass spectrometers may dispose different number of mass analyzers, and could use different ionization method such as electrospray, ionspray, matrix assisted laser desorption ionization (MALDI) to name the most frequently used method to analyze proteomics samples. In label-free LC–MS proteomics experiments, there are two types of widely used mass spectrometry data. The first data type contains mass spectra obtained with one mass analyzers and is referred to as single stage mass spectrometry data (MS-1) in the literature. The second data type is heterogeneous and contains cyclic series of MS-1 and precursor ion fragmented spectra (MS/MS). Each cycle begins with MS-1 spectra, then it is followed by a defined number (generally 1–10) of MS/MS spectra obtained from the most abundant ions of the MS-1 spectra. This acquisition mode is referred to as data dependent acquisition (DDA) and abbreviated as DDA MS/MS data. The reader is referred to dedicated books [38,40,41] and reviews [39,42,43] for further reading on the main character-